System-Level Heterogeneity with Intel® Xeon Phi™ Processors

Estela Suarez

Jülich Supercomputing Centre

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreements 287530 (DEEP) and 610476 (DEEP-ER).

Collaborative R&D in DEEP & DEEP-ER

European Union Exascale projects
- 20 partners
- Total budget: 28.3 M€
- EU-funding: 14.5 M€
- Combined term: 5 years

Visit us @ ISC 2016, Frankfurt (Germany)
June 19 – 23, 2016
Booth #1340

www.deep-project.eu www.deep-er.eu

DEEP/-ER Approach

Co-design between applications, system SW and HW

- Application and operational requirements do shape system architecture
- HW provides performance, scalability and energy efficiency
- System SW enables applications to leverage HW potential

Objectives

- Deliver highest scalability and workload performance
- Provide leading energy efficiency (energy per result)
- Offer familiar, easy to use programming environment and APIs based on standards
- Ensure sustainability of system architecture and SW

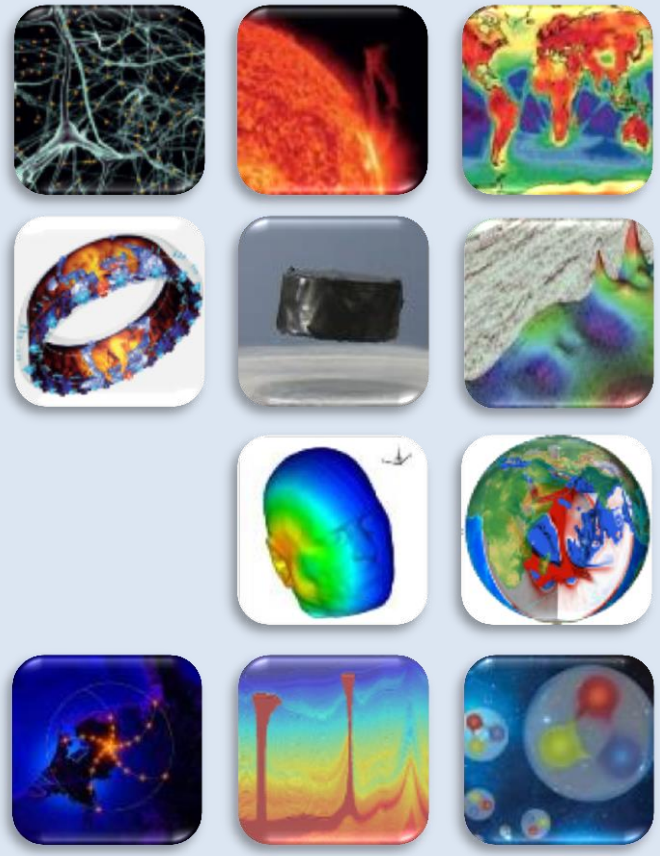
Design elements

- Exploit benefits of processor heterogeneity
- Leverage technology advances in storage-class memory and interconnects

Co-Design Applications

DEEP + DEEP-ER applications

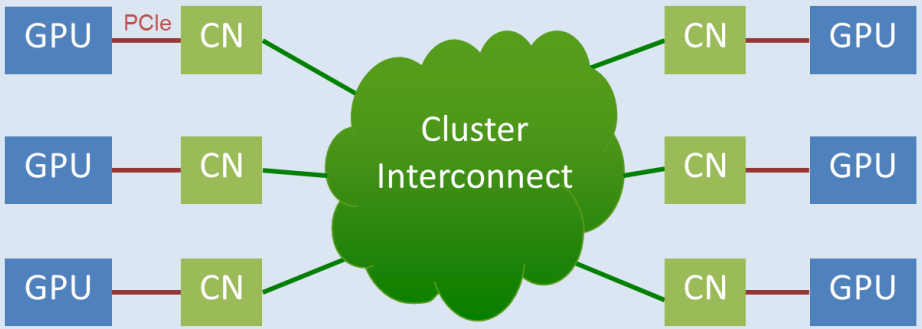
- Brain simulation (EPFL)
- Space weather simulation (KULeuven)
- Climate simulation (CYI)
- Computational fluid engineering (CERFACS)
- High temperature superconductivity (CINECA)
- Seismic imaging (CGG)
- Human exposure to electromagnetic fields (INRIA)
- Geoscience (BADW-LRZ)
- Radio astronomy (Astron)
- Oil exploration (BSC)
- Lattice QCD (UREG)



Goals

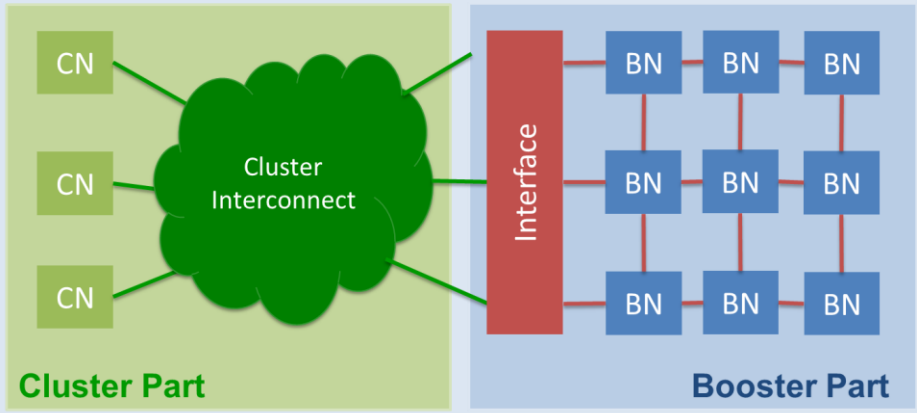
- Drive co-design cycle
- Evaluate and validate system prototypes

System-Level Heterogeneity



Accelerated Cluster

- Fixed, static ratio and assignment of accelerators to CPUs
- Static management of resources
- Accelerators do not act autonomously
- General-purpose Cluster interconnect
- Programming via local offload interfaces (OpenCL, CUDA, CELO, OpenACC, …)



Cluster-Booster Architecture

- No fixed ratio or assignment between resources (Multicore & Manycore nodes)
- Dynamic management and association of resources
- High-throughput network in the Booster
- Programming via MPI and “global” tasking interfaces

DEEP System Architecture

Cluster part

– High single thread & complex code performance → Intel® Xeon® processors

– High any-to-any connection performance & infrastructure integration → standard switched HPC fabric (InfiniBand™)

Booster part

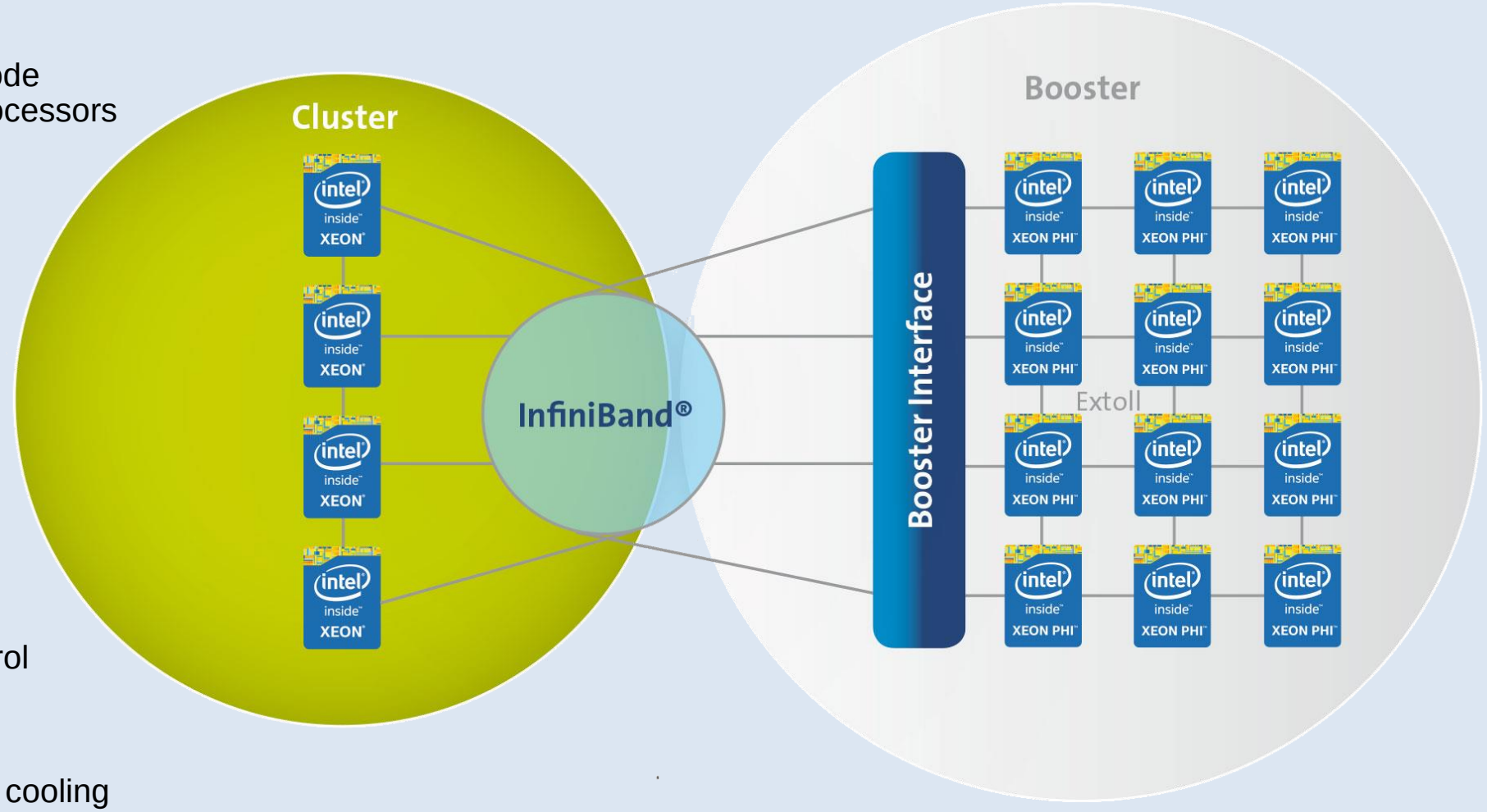
– High throughput, autonomous operation → Intel Xeon Phi coprocessor (codenamed “Knights Corner”)

– Need for low latency, spatial application structures → 3D Torus direct-connected network (EXTOLL)

– Network bridging and KNC control → Booster Interface layer

Both parts

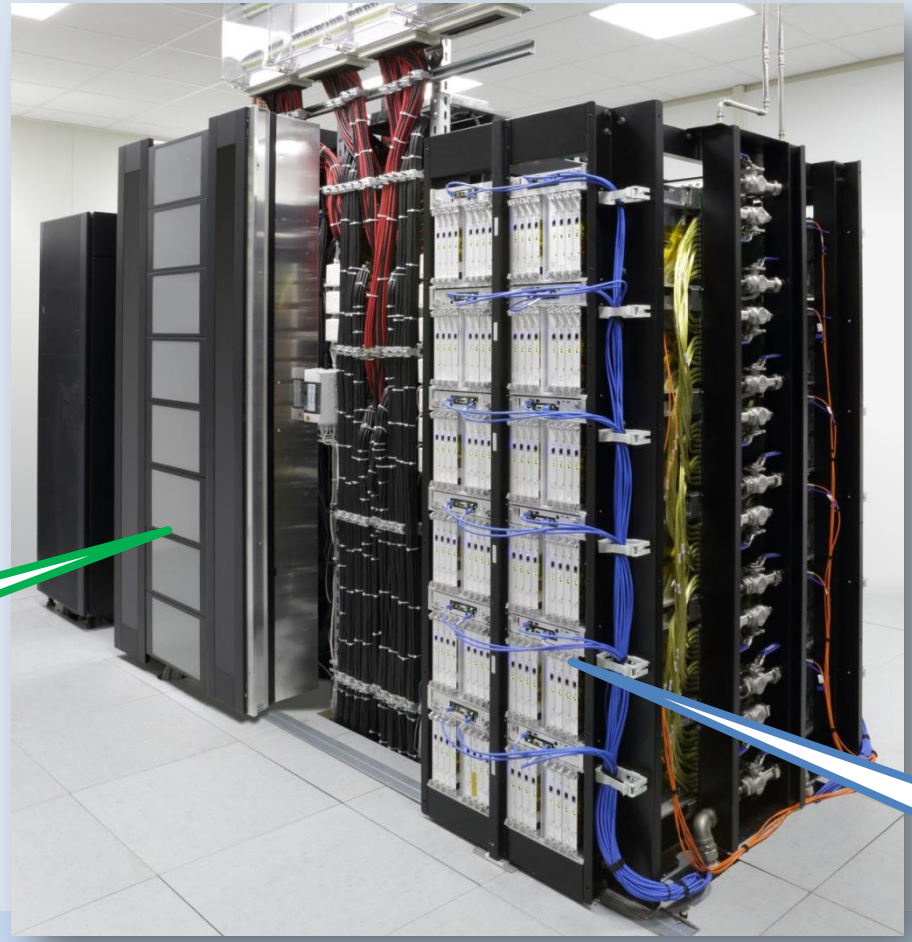
– Efficiency needs → use of liquid cooling & dense packaging



DEEP Prototype Systems

Eurotech Aurora Prototype

Cluster part
- 128 dual-socket Intel Xeon E5-2403 nodes
- QDR InfiniBand™
- Eurotech Aurora liquid cooling & packaging

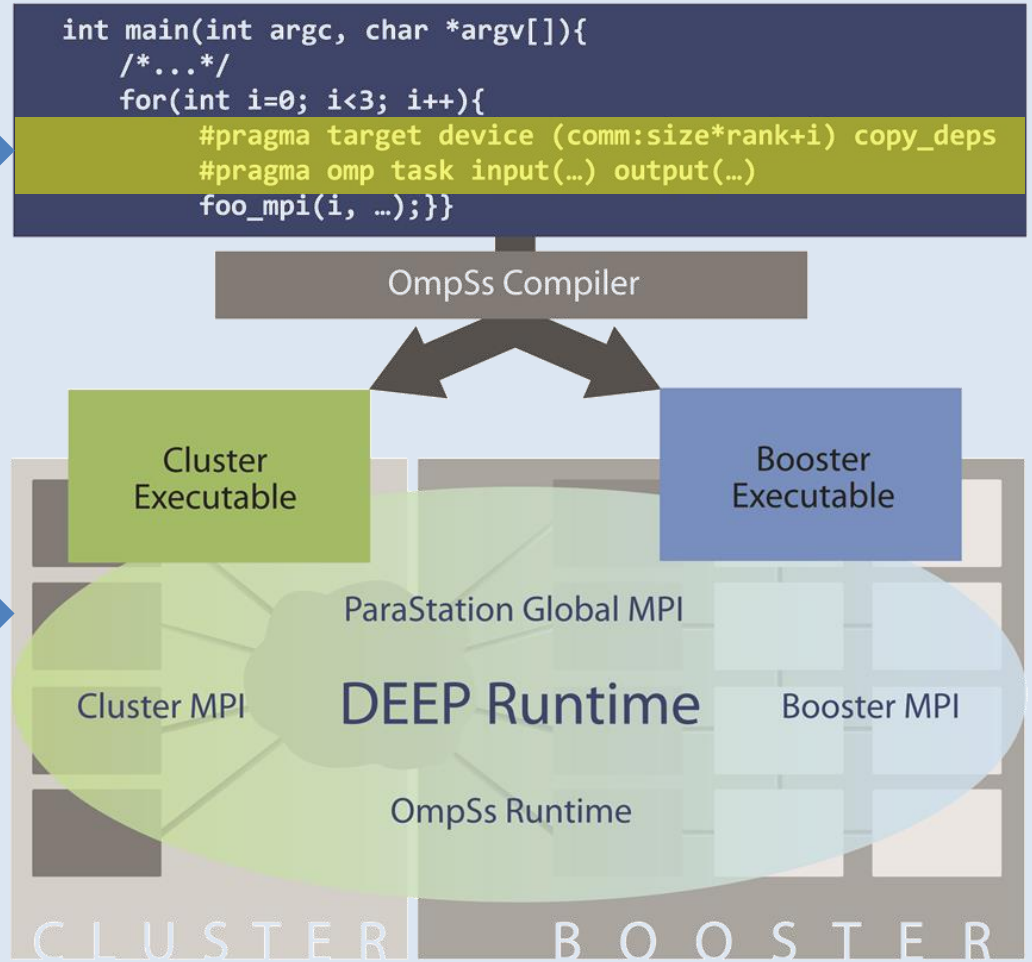


Booster part
- 384 Intel Xeon Phi 7120X nodes
- FPGA implementation of EXTOLL interconnect
- 24 Booster interface nodes with Intel Xeon processor
- Eurotech Aurora liquid cooling & packaging

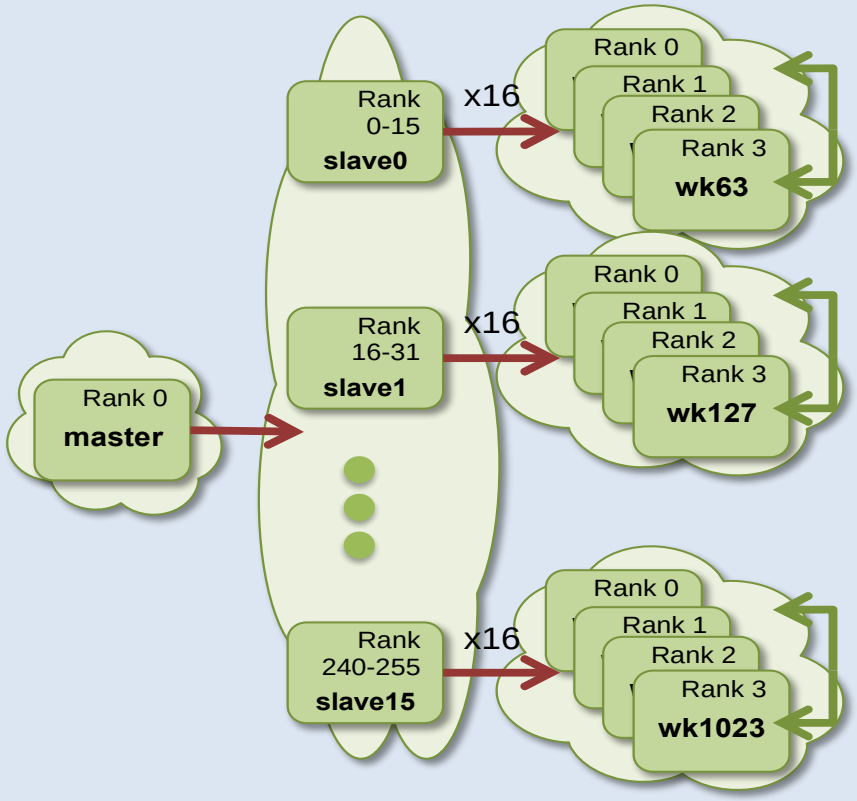
DEEP/DEEP-ER Programming Model

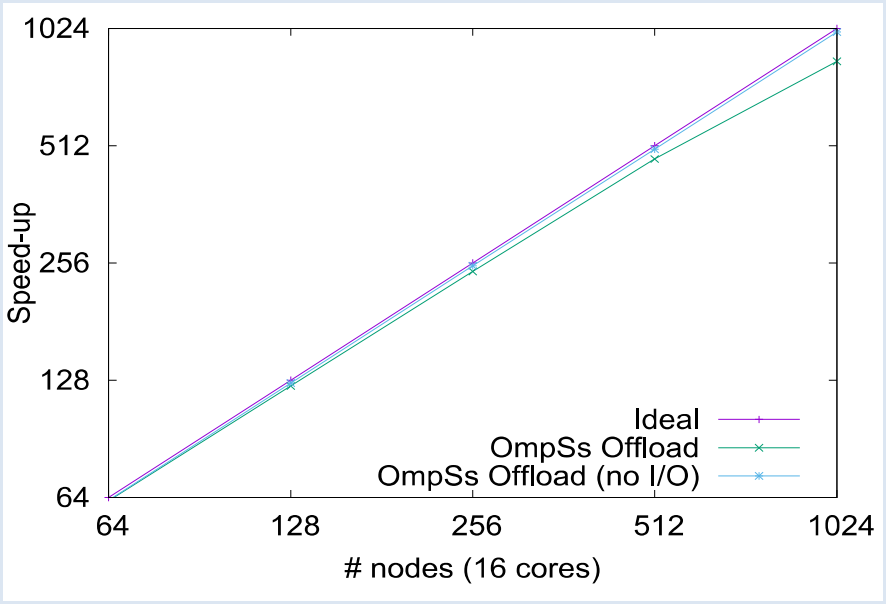
Task-based OmpSs programming model
- Pragma based, emphasizes ease of use
- Efficient communication across the whole system
- Dynamic spawning of massively parallel tasks in both parts

ParaStation Global MPI layer
- Expert-level programming
- Efficient communication across the whole system
- Dynamic process spawning an control in both directions



Massively Parallel Tasks in OmpSs

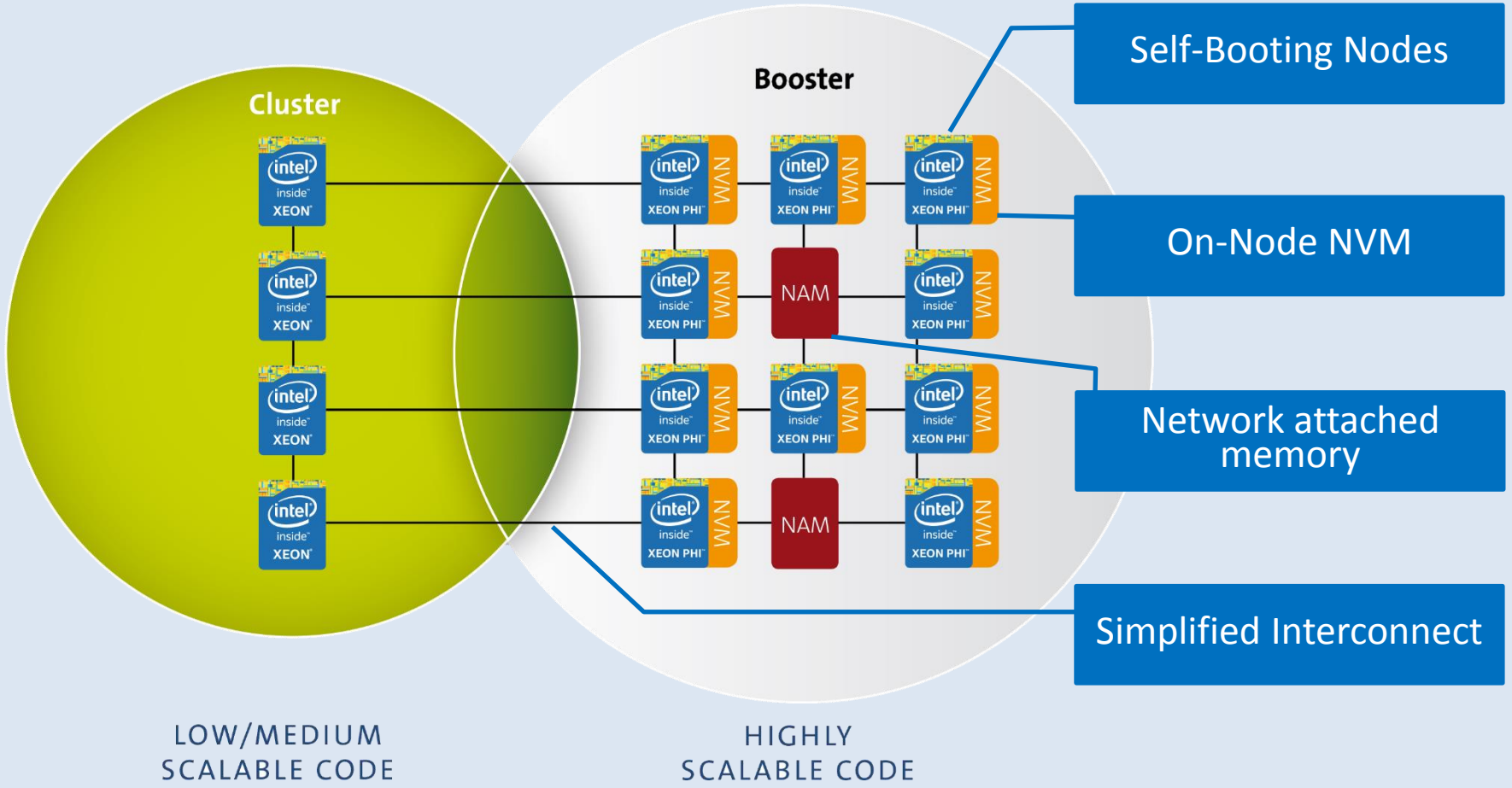




Measurements for BSC FWI (full waveform inversion) code

Published in: *Sainz, F, Bellón, J, Beltran, V, Labarta, J, "Collective Offload for Heterogeneous Clusters"*, IEEE 22nd International Conference on High Performance Computing (HiPC), 2015

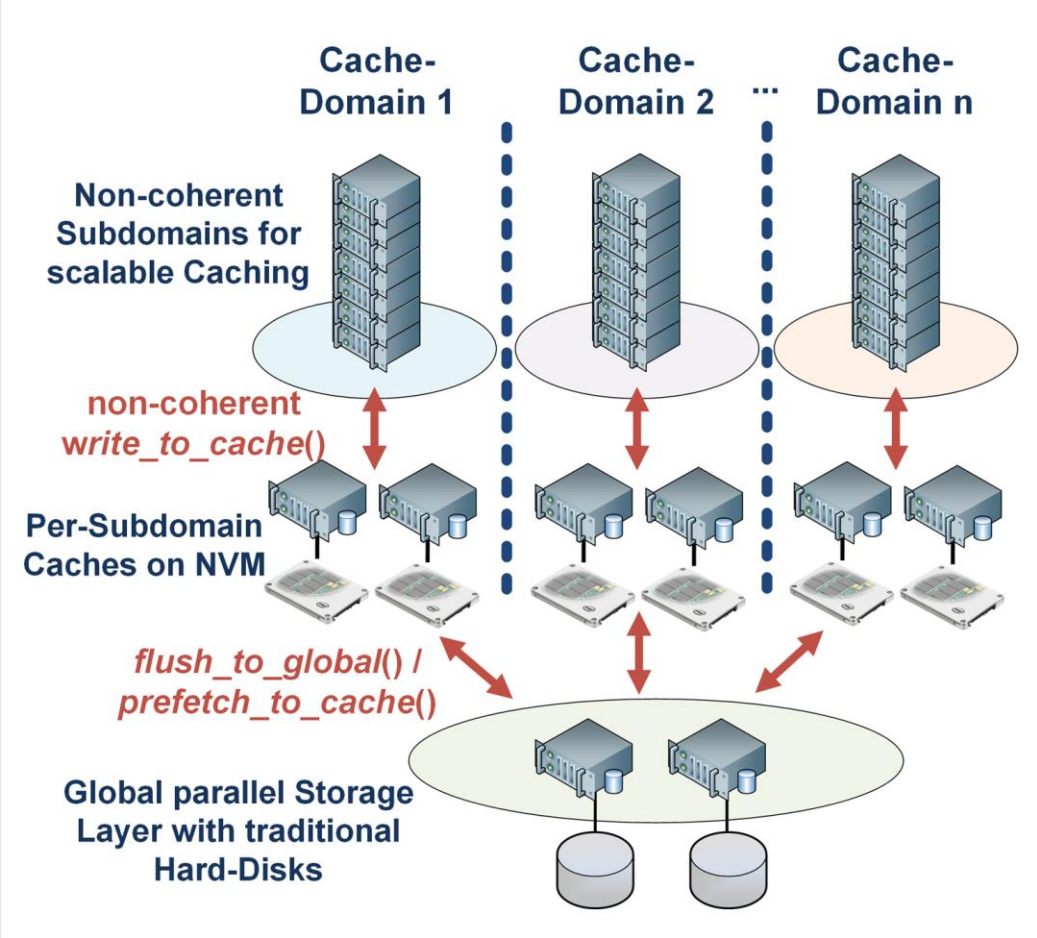
From DEEP to DEEP-ER



DEEP-ER Scalable I/O

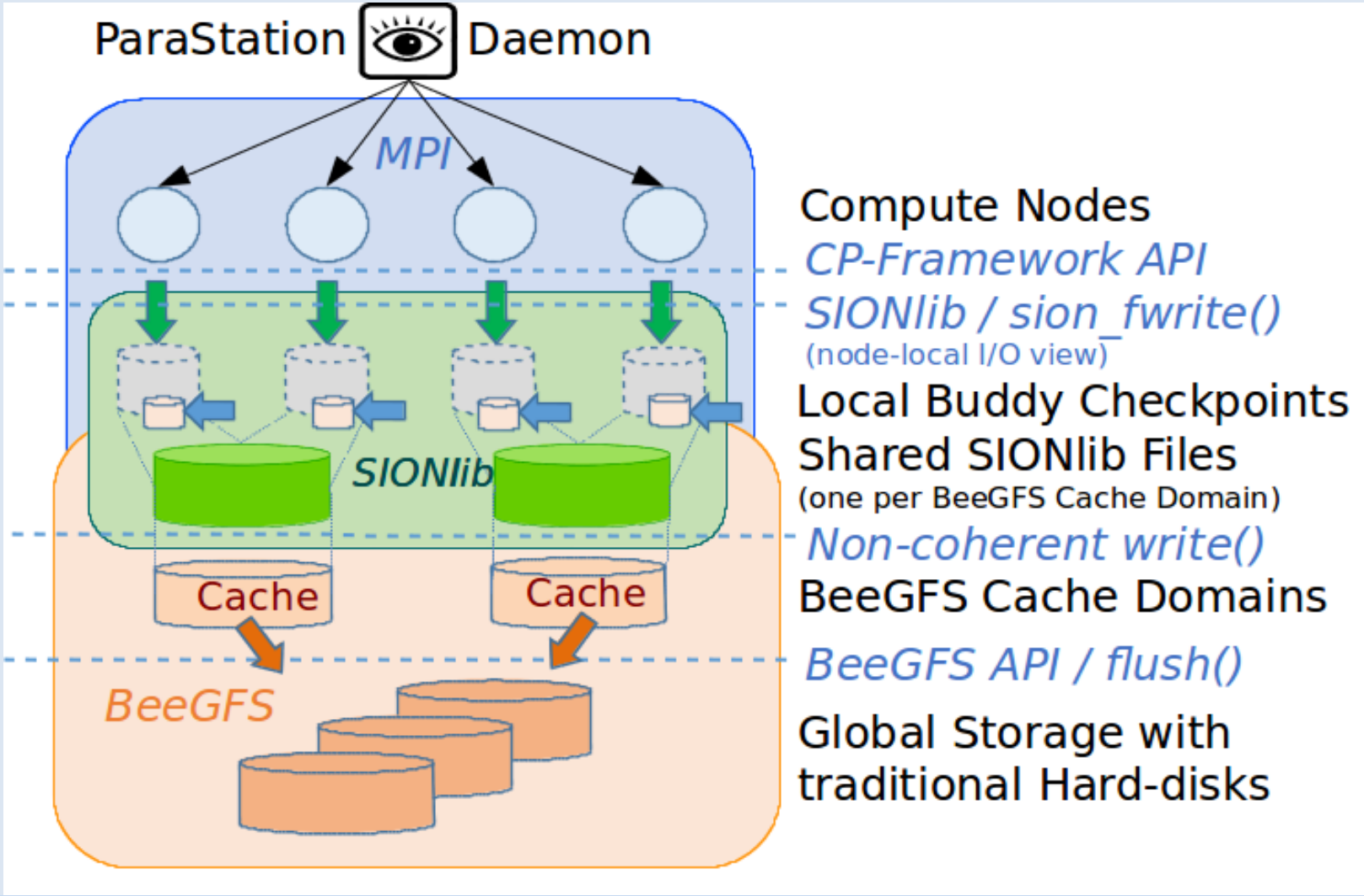
Leverage presence of fast local NVM storage

- Scalable caching of read/write data close to requesting node
- Prefetching stages read data into caches
- Write-back scheme saves data to permanent storage
- Synchronous (done) and asynchronous (WIP) versions/APIs

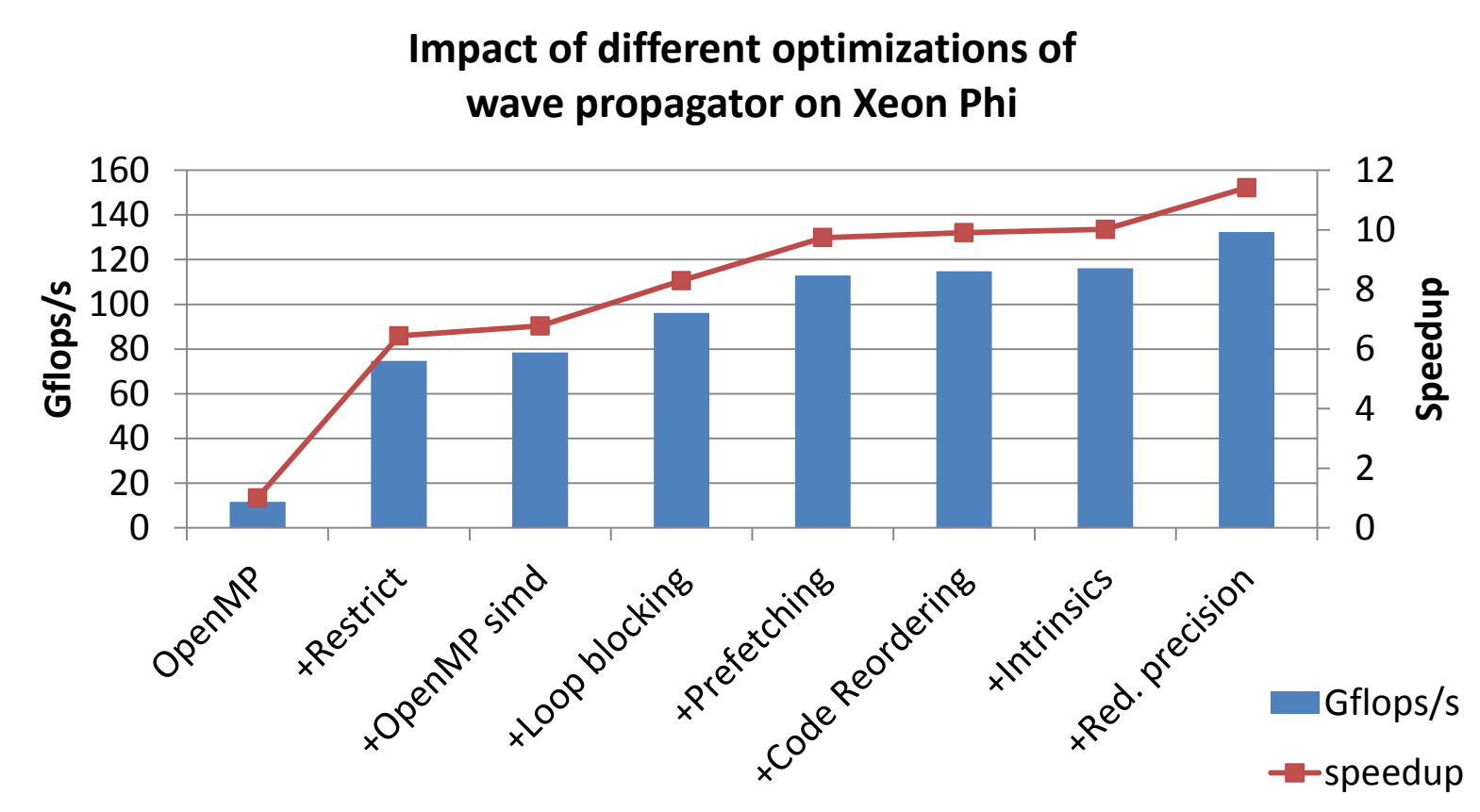


BeeGFS® developed by Fraunhofer

DEEP-ER Resiliency Scheme



BSC Full Waveform Inversion Results



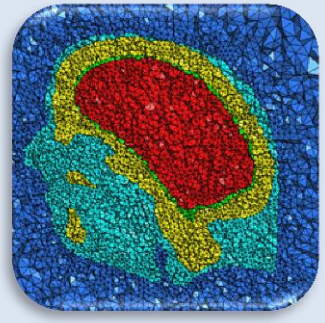
Using 60 cores per Xeon Phi coprocessor node with 180 threads

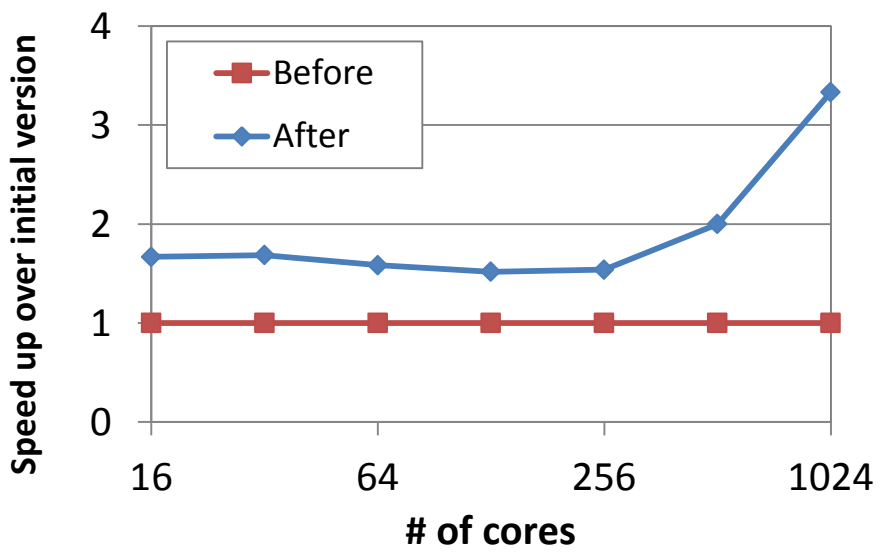
INRIA MAXW-DGTD Results

Improvements applied below:

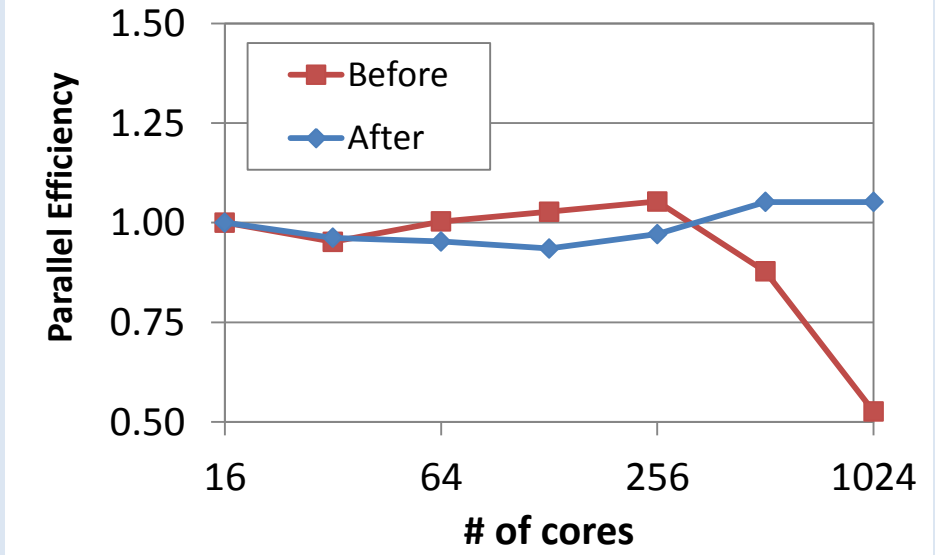
- Non-blocking communication
- Renumbering scheme
- Vectorisation and locality

Setup:
- Human head
- DEEP Cluster
- Mesh: 1.8 million cells
- 16 processes per node
- Pure MPI.
- P1 approximation.





Performance improvement up to 3.3x



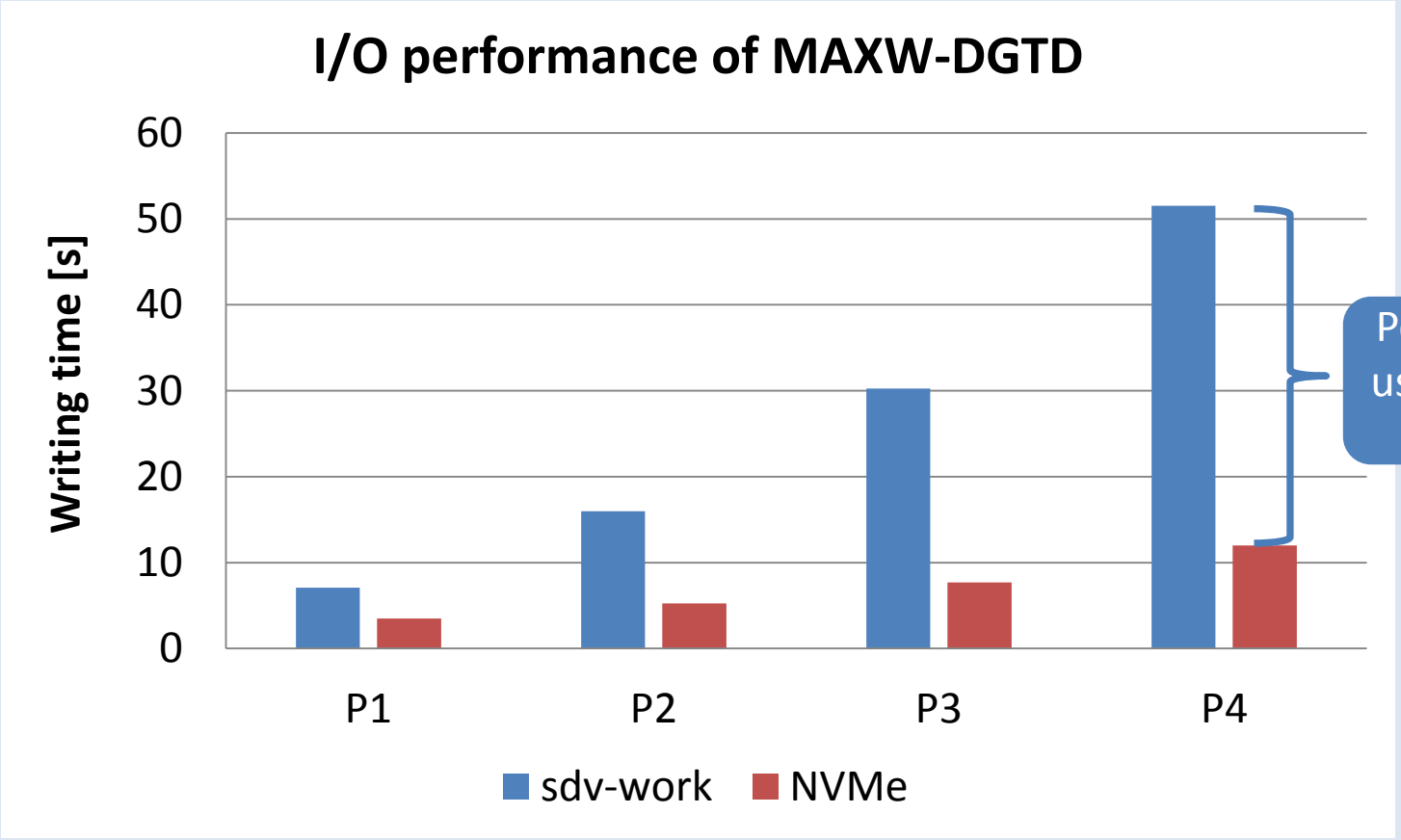
Almost perfect parallel efficiency now

*Leger R., Alvarez Mallon D., Duran A., Lanteri S., "Assessing the DEEP-ER Cluster/Booster Architecture with a fine-element type solver for bioelectromagnetics", Submitted to PARCO2015, Contribution ID: 25. www.parco2015.org

INRIA MAXW-DGTD I/O Results

Inria: Assessment of Human exposure to EM fields

24 MPI processes, 1 thread per process



Increasing model precision P1<P2<P3<P4

Partners

Fraunhofer ITWM

CGG
Passion for Geoscience

CINECA

The Cyprus Institute

EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

ASTRON

CERFACS

German Research School for Simulation Sciences

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EUROTECH
Imagine. Build. Succeed.

intel

Inria
INVENTORS FOR THE DIGITAL WORLD

ParTec Cluster Competence Center

KU LEUVEN

Mellanox TECHNOLOGIES

UR Universität Regensburg

SEAGATE

JÜLICH FORSCHUNGSZENTRUM

lrz

Ruprecht-Karls-Universität Heidelberg